DOCUMENT RESUME

ED 451 182 SP 039 865

AUTHOR Salzman, Stephanie A.; Denner, Peter R.; Bangert, Arthur W.; Harris, Larry B.

TITLE Connecting Teacher Performance to the Learning of All Students: Ethical Dimensions of Shared Responsibility.

SPONS AGENCY J.A. and Kathryn Albertson Foundation, Inc., Boise, ID.
PUB DATE 2001-03-00
NOTE 50p.; Paper presented at the Annual Meeting of the American Association of Colleges for Teacher Education (53rd, Dallas, TX, March 1-4, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Standards; Elementary Secondary Education; Higher Education; Job Performance; Performance Tests; Preservice Teacher Education; *State Standards; Student Teachers; *Teacher Competencies; *Teaching Skills; *Work Sample Tests

ABSTRACT

This paper presents the results of a benchmarking, validity, and generalizability study of the use of teacher work samples to assess the ability of preservice and inservice teachers to meet program and state teaching standards and to impact the learning of the students they teach. The assessment approach builds upon the Teacher Work Sample Methodology of Western Oregon University. A major goal of the study was to identify "benchmarks" or exemplars of performance along the full developmental continuum from beginning to expert teaching by having sample groups of early interns, student teaching interns, experienced teachers, and National Board Certified teachers complete teacher work samples. The study also examined whether work samples could be feasibly and equitably administered and scored with sufficient reliability to warrant their use for high-stakes decisions about the effectiveness of teaching performance. Results of the study show initial support for teacher work sample assessment as a way to provide valid and credible evidence connecting teaching performance to student learning. (Contains 18 references.) (Author/SM)

Reproductions supplied by EDRS are the best that can be made
from the original document.

SP

ED 451 182

Connecting Teacher Performance to the Learning of All Students: Ethical Dimensions of Shared Responsibility

Stephanie A. Salzman

Peter R. Denner

Arthur W. Bangert

Larry B. Harris

Idaho State University

Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, Dallas, Texas, March 1-3, 2001.

Author Note: This study was partially supported by a grant from the J. A. & Kathryn Albertson Foundation. Send correspondence to Stephanie Salzman, College of Education, Idaho State University, Box 8059, Pocatello, ID 83209 or salzstep@isu.edu.

SP039865

BEST COPY AVAILABLE

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

S. A. Salzman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U S DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Abstract

This paper presents the results of a benchmarking, validity, and generalizability study of the use of Teacher Work Samples to assess the ability of preservice and inservice teachers to meet program and state teaching standards and to impact the learning of the students they teach. Our assessment approach builds upon the "Teacher Work Sample Methodology" of Western Oregon University (Schalock, 1998; Schalock, Cowart, & Staebler, 1993). A major goal of our study was to identify "benchmarks" or exemplars of performances along the full developmental continuum from beginning to expert teaching by having sample groups of early interns, student teaching interns, experienced teachers, and National Board Certified teachers complete teacher work samples. We also examined whether work samples could be feasibly and equitably administered and scored with sufficient reliability to warrant their use for high-stakes decisions about the effectiveness of teaching performance. Results of the study show initial support for teacher work sample assessment as a way to provide valid and credible evidence connecting teaching performance to student learning.

Connecting Teacher Performance to the Learning of All Students: Ethical Dimensions of Shared Responsibility

The National Commission on Teaching and America's Future (1996) through its report, titled *What Matters Most*, articulated an imperative to establish high and rigorous standards for what teachers should know and be able to do and to advance related education reforms for the purpose of improving student learning. Consistent with this call to action, the National Council for Accreditation of Teacher Education (NCATE, 2000) established new accreditation standards requiring documentation of the impact of program candidates and graduates on the learning of the students they teach. To effectively respond to these mandates, institutions that prepare teachers must set higher standards for teacher candidates and then provide in-depth learning experiences that enable candidates to meet the standards. Concomitantly, teacher education institutions must develop and implement assessment systems that yield defensible and credible evidence regarding candidates' ability to meet these standards and impact PK-12 student learning.

In responding to these mandates, teacher education programs are faced not only with an urgent need to devise assessments that supply credible evidence of candidate performance but they are also faced with the ethical imperative to institute assessment practices that meet technical standards for sound professional practice (American Psychological Association, 1985). Technical standards cover such issues as the quality of the assessment instruments, their propriety for the specific purposes for which they are used, including evidence of both validity and reliability, and the reasonableness of inferences based on their results. The latter is particularly important when assessments

are used to qualify students for teaching certification. Sound practice also means teacher education programs must develop guidelines for administering assessments and for protecting the rights of candidates, including informed consent and confidentiality in reporting and maintaining performance records. Because teacher education programs are largely a joint enterprise of teacher education faculty, faculty from colleges of arts and sciences, and practicing educators, responding to these mandates also necessitates a shared ethical responsibility.

This study addresses the development of teacher assessments that examine student learning as a function of teachers' work, while at the same time providing supporting evidence of candidates' ability to meet program and state standards. Our assessment approach is built upon the Teacher Work Sample Methodology (TWSM) of Western Oregon University (Schalock, 1998; Schalock, Cowart, & Staebler, 1993; Schalock, Schalock, & Girod, 1997). Teacher work samples are complex performance assessments in which teacher education candidates (or practicing teachers) are asked to document their teaching of an actual set of lessons. The documentation includes planning for instruction, the design of an instructional sequence usually covering at least four weeks of instruction, a plan for the assessment of learning both pre- and post instruction, demonstration and analysis of the impact of instruction on student learning, and reflection upon the success of the instructional unit. An important aspect is the requirement for teachers to demonstrate the consequences and results of their teaching in terms of its impact on student learning. Thus, the use of Teacher Work Sample Methodology holds great promise as an accountability tool for providing credible evidence of the impact of program candidates and teacher education graduates on the learning of the students they teach (for further discussion see Schalock, 1998).

While we agree Teacher Work Sample Methodology (Schalock, 1998) holds great promise for responding to the mandates for teacher education program accountability, we found early in our implementation of the approach that ethical considerations related to sound assessment practice needed to be addressed. Moreover, critics (Airasian, 1997; Darling-Hammond, 1997; Stufflebeam, 1997) have suggested important issues of reliability and validity are as yet unresolved. Most important among these technical and ethical issues is whether TWSM produces assessments of teacher performance of sufficient validity, freedom from bias, and reliability to warrant their use in high stakes decisions about teaching performance. In particular, it is important to establish the validity of the work sample assessments and the reliability of the ratings when the scoring rubrics are used by non-partisan raters (Popham, 1997). A further ethical consideration of particular interest to us is the extent to which the teacher work sample assessments authentically represent teachers' work.

To address these technical and ethical considerations, teacher work samples must be built upon clearly articulated standards, expert raters must have focused training, and raters must apply common standards-based criteria to judge performance. As we adapted Western Oregon's Teacher Work Sample Methodology in our undergraduate teacher preparation context, we quickly found that in order to address these ethical dimensions of assessment, we had to revise the approach in a number of aspects, including the way the work samples are structured and scored. We also had to determine how we were going to develop credible evidence of validity and scoring reliability.

A major aim of our benchmarking study was to support the validity of our work sample assessments for the purpose of documenting candidates' ability to meet

program and state teaching standards targeted by the assessment. A second goal was to establish models of acceptable and unacceptable work sample performance by identifying benchmarks or exemplars of performances along the full developmental continuum from beginning to expert teaching by having sample groups of early interns, student teaching interns, experienced teachers, and National Board Certified teachers complete teacher work samples. A third goal of our study was to determine whether work samples could be feasibly and equitably administered and scored with sufficient inter-rater reliability to warrant their use in high-stakes decisions about the effectiveness of teaching performance. As a final goal, we sought further support for the validity of work sample assessments for providing credible evidence of the impact of teaching performance on student learning.

To obtain a range of work samples for our benchmarking study, we solicited the involvement of teacher education candidates, experienced teachers, and highly accomplished National Board Certified Teachers to complete teacher work samples according to our guidelines. The teacher education candidates completed work samples as part of their program and course requirements. They gave informed consent for the use of their work samples in this study. The teachers were volunteers who completed teacher work samples because of their belief in their shared responsibility for developing credible teacher education program assessments. Many of them also volunteered because they responded to the moral imperative to connect their performance to the learning of their students. This involvement of practicing teachers enabled us to compare performances along the full continuum of professional development from novice to expert.

Adapting benchmarking procedures developed by the National Board for Professional Teaching Standards (A. Harmon, personal communication, June 1, 2000), we then recruited well qualified expert raters, including National Board Certified Teachers, to serve as judges for our benchmarking activities. In addition to both holistic and analytic ratings of the work samples, the benchmarking activities resulted in the expert raters identifying exemplars at each level of performance on a developmental continuum from beginning to exemplary level. We envisioned the benchmarking study as fulfilling the dual purposes of establishing the validity and reliability of the teacher work sample methodology and providing training for the individuals who would later share responsibility for the teacher education program assessment process.

Methods

Teacher Work Sample Guidelines and Scoring Rubrics

As our first step in developing our work sample assessments, we worked collaboratively with our professional community to examine the Idaho Core Teacher Standards (Idaho State Board of Education, 2000) and our institutional Beginning Teacher Core Standards (College of Education, 1995) to set the targeted standards for the teacher work sample (see Appendix A). Once the targeted standards were set, we defined indicators of the standards that our professional community agreed provided the evidence of performance one would look for to evaluate whether or not the targeted standards were met. The generation of the targeted standards and indicators involved widespread discussion with opportunities for input from our constituencies and culminated in an institutional decision to support the targeted standards and indicators as the basis for making decisions regarding candidate performance.

Using the standards and indicators as a framework, we then developed work sample tasks with accompanying directions to elicit the performances we sought to assess. The directions took the form of a set of Teacher Work Sample Guidelines (see Appendix B) designed to take each candidate step-by-step through the development of the work sample tasks. During the development of the guidelines, we took extra care to ensure absolute alignment between the standards and indicators and the components of the work sample. While the general framework for our teacher work sample tasks closely resembles that of Western Oregon University (Schalock, Cowart, & Staebler, 1993), we included significant revisions to reflect our targeted program standards. Our teacher work sample tasks require candidates to develop a written product that includes the following components: (1) a description and analysis of the learning-teaching context, (2) achievement targets for the instructional sequence, (3) an assessment plan, (4) plans for an instructional sequence comprised of at least six related learning activities aligned to the achievement targets to be taught over a four-week time period, (5) analysis of student learning, and (6) evaluation and reflection on the success of the instructional sequence with regard to student learning and future practice. In addition to specific directions for the development of each of these components of the work sample, the guidelines also included a template for the format for each learning activity plan (see Appendix B).

Using the targeted standards and indicators, we also developed an analytic scoring rubric (see Appendix C) that provides specific feedback to candidates regarding their performance on each of the targeted standards. The analytic scoring rubric lists the targeted standards with a description of the indicators for each standard that become the criteria for judging performance relative to the standard. Each of the six targeted

standards for the teacher work sample is rated on a 3-point scale: 0 = Standard Not Met; 1 = Standard Partially Met; and 2 = Standard Met.

While the analytic scoring rubric provides specific feedback to candidates relative to each of the standards, we found we needed an additional scoring rubric that would enable us to make a holistic judgment regarding the total performance of our teacher education candidates on the teacher work sample assessment. The holistic scoring approach reflects the complex nature of teaching and avoids the error of disaggregating the performance and, as a result, diminishing authenticity or realism. With the assistance of A. Harmon (personal communication, June 19, 2000) from the National Board for Professional Teaching Standards, we designed a holistic scoring rubric that categories the total performance on a developmental continuum: 1 = Beginning; 2 = Developing; 3 = Proficient; and 4 = Exemplary (see Appendix D). The holistic score defines the level of performance in terms of an overall judgment of the degree to which the teacher work sample provides evidence of meeting all six of the targeted standards.

Benchmarking Participants

To obtain a representative range of performances on the teacher work samples, we not only required our junior-level (early internship) and senior-level (student teaching internship) teacher education candidates to complete work samples, but also recruited practicing teachers, including National Board Certified teachers, to develop work samples. This involvement of candidates, student teachers, experienced teachers, and highly accomplished National Board Certified teachers helped to ensure the identification of exemplars of performances along the full continuum of professional development from novice to expert. A set of n = 132 work samples were collected. Of these, 54 were from junior level practicum students, 44 from senior level students

completing their student teaching internship, 30 from classroom teachers, and 4 from National Board Certified teachers. The work samples represented a range of subject areas, including 33 English/Language Arts, 1 Communication, 3 Foreign Language, 9 Health, 16 Mathematics, 5 Professional/Technical, 5 Physical Education, 29 Science, 26 Social Studies, and 5 Visual/Performing Arts. All grade levels from K to 12 were represented in the set of work samples. There were 6 kindergarten work samples, 12 first grade, 21 second grade, 12 third grade, 8 fourth grade, 9 fifth grade, 7 sixth grade, 10 seventh grade, 16 eighth grade, 10 ninth grade, 3 tenth grade, 12 eleventh grade, and 6 twelfth grade.

Production and Collection of Work Samples

One of the most important steps in the use of the teacher work sample approach to assessment is communication of the tasks to be performed to the people developing the work samples. Because of its complexity, the development of a teacher work sample requires extensive guidelines and directions for its completion. To aid clear communication of the tasks, all participants received a document titled Teacher Work Sample Guidelines for Preparation (see Appendix B), which delineated the required components and the necessary steps for preparing them.

Because the guidelines are complex, and the development of a work sample demands the application of broad knowledge and multiple skills and strategies required for an authentic representation of the teaching process, we have developed an approach through which our teacher candidates are "scaffolded" during the development of their first teacher work sample. All of our candidates complete two teacher work samples during our teacher education program. The first work sample is completed as a requirement for a junior-level course that includes a half-time internship

in a PK-12 classroom. As these junior-level teacher education candidates develop their first work sample, they are given intensive mentoring and instruction in the knowledge and skills required for its successful completion. The second teacher work sample is completed during the senior-level student teaching internship. Unlike the first work sample, the second work sample is completed independently by the candidate.

The practicing teachers who participated in this benchmarking study received directions and support via a two-credit professional development course taught by a College of Education faculty member and an elementary school principal. The course did not provide the teachers with the same level of mentoring and instruction received by the junior-level teacher education students. It was assumed the practicing teachers possessed the knowledge and skills necessary to complete the work samples. Instead, support focused on the expectations of the requirements for the work samples and on answering the questions the teachers had related to the specifics of the work sample components and how each component should be documented. The two professional development credits served mainly as compensation for the time the teachers devoted to the development and submission of their work samples. The course credits are not indicative of the amount of assistance the teachers received. The teachers completed their work samples on their own in a manner similar to our senior-level student teaching interns.

Panel of Expert Raters

Because our teacher work sample assessment process involves cooperating teachers and arts and sciences faculty in assessing candidate performance relative to our program standards, we included representatives of these constituencies in the benchmarking study as expert raters. The public school representatives on the team of

raters included 8 teachers and 1 principal. Five of the public school representatives worked in elementary schools and 4 in junior high schools. Eight were women and 1 was a man. Five of the teachers held a bachelor's degree (plus credits), 3 of the teachers and the principal held master's degrees. Together these raters had a median of 18 years (ranging from 11 to 30 years) of public school teaching experience. Three of the teachers were National Board Certified. The faculty representatives on the team of raters consisted of 5 Division of Teacher Education faculty members, 1 College of Arts and Sciences faculty member, and 1 part-time supervisor of student teaching interns. Five of the faculty members were women and two were men. Five faculty members held a doctoral degree, while two of the faculty members held a master's degree (plus credits). The faculty members had a median of 9 years of public school teaching experience (ranging from 0 to 26 years), and a median of 15 years of college teaching experience (ranging from 5 to 22 years).

Procedures

The benchmarking study was comprised of two consecutive one-day sessions. The first day was spent on training for uncovering potential scoring bias and identifying exemplars at each level of the holistic scoring rubric. At the end of the first day, we also gathered content validity data. On the second day, the expert raters scored the exemplar teacher work samples using the analytic scoring rubric.

Because of potential scoring bias due to personal preferences regarding good teaching, prior to beginning benchmarking activities, we conducted training targeted toward uncovering personal biases. As the first step in this training, the expert raters were directed to list characteristics of excellent teachers and characteristics of very poor teachers. After the lists were completed and small-group discussions were conducted,

the expert raters compared the characteristics they wrote on their personal lists to the standards (see Appendix A) targeted in the work sample. Those characteristics of either excellent teachers or poor teachers that did not appear in the standards were recorded by each judge on his or her "Hit List of Personal Biases." These hit lists were used by the expert raters during benchmarking and scoring as constant reminders to focus on the standards as the sole lens for scoring the teacher work samples.

The next step in preparing the expert raters for scoring the teacher work samples consisted of reviewing general guidelines for scoring. These guidelines addressed such issues as security, halo and pitchfork effects in scoring, and the importance of focusing on evidence found throughout the work sample. As a group, the expert raters were then taken through a review of the Teacher Work Sample Standards and Indicators (see Appendix A) and the level of performances defined in the holistic scoring rubric.

The first goal of the benchmarking activity was to identify exemplars of performances at each level of the holistic scoring rubric. The raters were divided into groups. Each group then performed a "quick read" of approximately 20% of the 132 work samples. After this, each group then reached consensus on the holistic score category and placed the work sample in one of four piles representing the four levels of the scoring rubric. In the afternoon, the work samples within a category were then scored a second time by a different group of raters and, after discussion, two or three exemplars of performance at that level were identified. This resulted in the establishment of three sets of 10 exemplars consisting of 2 exemplars at the Beginning level, 3 exemplars at the Developing level, 3 exemplars at the Proficient level, and 2 exemplars at the Exemplary level. Within levels, the exemplars were randomly assigned to the three sets.

Following holistic scoring of the work samples and the identification of the three sets of exemplars, we used the same expert raters to gather validity evidence. We applied Linda Crocker's (1997) methodology for performing content judgments of performance assessment exercises and scoring rubrics. The criteria used for judging the teacher work sample as an assessment exercise included criticality of the behavior, frequency of the behavior in job performance, and realism of the teacher work sample as a simulation of actual classroom performance. The process for making content judgments regarding the scoring rubric involved matching the elements of the exercise and the scoring rubric to the assessment domain (i.e., the targeted standards - see Appendix A). In addition, the raters matched the elements of the teacher work sample and the scoring rubric to the Idaho Core Teacher Standards (Idaho State Board of Education, 2000).

The following day, the same raters returned. After the directions for the analytic scoring rubric were explained, each of the raters was randomly assigned to analytically score one of the 3 sets of 10 work samples. Thus, 5 raters each scored the same 10 work samples contained in one of the three sets. Each rater continued to use her or his "Hit List of Personal Biases." The raters were exhorted to score the work samples on the basis of the standards and indicators contained in the analytic scoring rubric only. Each rater scored their assigned work samples independently.

Results

Holistic Scoring Method

Using the holistic scoring rubric, of the n = 132 work samples categorized by the expert raters, 25 (18.9%) were judged to be Beginning, 49 (37.1%) were judged to be Developing, 37 (28.0%) were judged to be Proficient, and 21 (15.9%) were judged to be

Exemplary. Surprisingly, there was no statistically significant association (at the $\alpha = .05$ level) between the holistic score categorizations and the source of the work samples (junior level interns, student teaching interns, teachers, or National Board Certified teachers), χ^2 (df = 9) = 15.76, p = .07. Happily for our benchmarking purposes, the results indicated that all levels of teaching proficiency were evidenced across our work samples in sufficient proportions for our raters to be able to choose several sets of exemplars. Importantly also, there was no statistically significant association found between the holistic score categories and the grade level of the work samples (elementary versus secondary), χ^2 (df = 3) = .66, p = .88, or subject area of the content of the work samples (English/Language Arts, Math, Science, Social Studies, or Other), χ^2 (df = 12) = 4.85, p = .96. This means the raters' judgments about teaching proficiency as evidenced by the work samples were not influenced by these factors.

Analytic Scoring Method

For the analytic scoring method we computed total score dependability coefficients for absolute decisions based on formulas provided by Crocker and Algina (1986) and Shavelson and Webb (1991). Table 1 presents the analysis of variance for the effect of rater for the three sets of teacher work samples. For all three sets, the effect of rater was not statistically significant at the $\alpha = .05$ level of significance. Table 2 presents the variance components used in the formulas for computing dependability for each of the three sets of work samples. Each set of work samples was scored by five different raters. The results yielded 5 rater coefficients of dependability for the three sets of work samples of .91, .88 and .94 respectively. These dependability coefficients are similar in interpretation to classical test theory's reliability coefficients.

Insert Table 1 about here

Insert Table 2 about here

Single Rater coefficients of dependability for absolute decisions for the three sets of work samples were computed to be .68, .60, and .75. Adjusting the number of raters included in the formula revealed an acceptable level of dependability of .75 to .86 for performance evaluations could be achieved with as few as two raters. These findings suggest work samples can be feasibly administered and scored with sufficient inter-rater reliability to make decisions regarding the quality of teaching performance. For our purposes, the above findings also showed that the average rating of the five raters of our three exemplar sets had sufficient dependability to be used as benchmark ratings for the training and calibrating of future raters.

Relationship Of Holistic to Analytic Scoring

The results of our study showed the two types of scoring, holistic and analytic, corroborated one another, while at the same time providing distinctive information about teaching performance. A single factor ANOVA using the unique sums of squares approach for unbalanced designs was conducted on the total analytic scores (averaged across the five raters) for the 30 work sample exemplars. The four holistic score categories served as the independent variable. The results revealed a statistically significant difference in total analytic scores received across the holistic scoring categories, F (3, 26) = 19.01, $p < .001$, MSE = 2.08. Post hoc mean comparisons using the

Tukey-Kramer procedure revealed a statistically significant difference ($p < .05$) between the analytic score means of the work samples categorized as Beginning level at $M = 5.00$ ($SD = 1.63$) and those categorized at higher levels. The means for the three other groups respectively were $M = 8.09$ ($SD = 1.39$) for Developing, $M = 10.16$ ($SD = 1.13$) for Proficient, and $M = 10.27$ ($SD = 1.73$) for Exemplary. In addition, the analytic score mean of the work samples categorized as Developing ($M = 8.09$) was found to be statistically significantly lower ($p < .05$) than the means of the work samples categorized as Proficient ($M = 10.16$) or categorized as Exemplary ($M = 10.27$). The latter two groups did not differ statistically. Hence, the four holistic scoring categories with the exception of the last two categories were distinguished by their average analytic ratings. The fact that the last two groups were not distinguished is an artifact of the analytic scoring method, which did not include a rating level beyond the level of standard met. Our analytic scoring procedure was not intended to distinguish exemplary from proficient performances and it did not do so.

Time Required to Score Work Samples

We also considered the amount of time necessary to score the work samples. Due to our two stage approach to holistically scoring the work samples, we were not able to track separately an exact time for the length of a typical holistic scoring. However, based on the total time it took for the teams to complete their holistic scoring of all of the work samples and the fact that each group scored approximately 20% of the work samples, we could estimate the time for holistically scoring a teacher work sample to be about 9 to 10 minutes.

Importantly, we were able to precisely measure the length of time it took to analytically score each of the work samples selected as exemplars. The average time for

scoring the n = 60 work samples was M = 13.5 minutes with a standard deviation of SD = 5.4 minutes. As expected, some raters took consistently longer to score their assigned work samples than others. Fortuitously, additional correlational analyses showed that scoring time was not correlated with total analytic scores for any of the three sets of work samples, r = .07, n = 50, p = .63 for Set 1, r = .18, n = 50, p = .20 for Set 2, and r = .11, n = 50, p = .46 for Set 3. These data demonstrate that the time it takes to reliably score teacher work samples is within a range that is realistic and practical. It should be noted, however, that these times were based on the analytic scoring of the work samples that were chosen as exemplars. Somewhat longer time might be required to analytically score work samples less exemplary of category membership and closer to the holistic category boundaries. This issue will be examined in our follow-up investigations.

Validity

To make content judgments regarding the validity of our teacher work sample assessment and scoring rubrics we applied the three criteria of realism, criticality, and frequency suggested by Crocker (1997) for judging the content representativeness of performance assessments and rubrics. The results are reported both in terms of our rationale supporting the adequacy and appropriateness of the matches among the elements of the work sample, the scoring rubrics, and the targeted assessment domain (i.e, the standards assessed by the work sample) and in terms of the empirical evidence supplied by the evaluative judgments of our panel of expert raters.

Requiring our teacher education candidates and practicing teachers to perform teaching tasks in actual public school classrooms speaks directly to the realistic nature of the teacher work sample assessment. Realism was supported by the fact that the

performance tasks were not simulations but actual lessons developed for and delivered to appropriate students in public school classrooms. Support for the realism of our teacher work sample assessments is also evidenced by the clear link between the richly detailed rubrics and the primary traits of proficient teacher performances reflected in the indicators of our targeted standards (see Appendix A). To support this, we had our expert raters evaluate the relationship between the work sample components, program standards and the actual work of teachers. All panel members agreed that the elements of the work sample, the scoring rubrics and the targeted standards were in alignment. Hence, our teacher work sample meets the criteria of a realistic assessment because it is a direct assessment consisting of open-ended activities that permit the use of multiple strategies for demonstrating application of knowledge and skills important to proficient teaching.

The panel of experts were also asked to judge whether the work samples measured knowledge and skills necessary for a beginning teacher. The results were 68.8% (n = 11) of the expert raters said "absolutely yes," 18.8% (n = 3) said "yes," while only 12.5% (n = 2) were "uncertain." We also asked the expert raters to assess the importance or criticality of the teaching behaviors that the teacher work samples required the candidates to demonstrate to actual teaching. The results yielded the same percentages, with 68.8% (n = 11) of the expert raters rating the teaching behaviors as "critical," 18.8% (n = 3) rating them as "important," and only 12.5% (n = 2) rating them as "somewhat important." None of the raters indicated the teaching behaviors were of little or no importance. These results support the criticality criteria for the content representativeness of the teacher work samples.

Next, we asked our panel of experts to indicate, using a scale of: 1 = Not at all; 2 = Implicitly; or 3 = Directly, the extent to which the tasks required for the teacher work sample reflected the Idaho Core Teacher Standards (Idaho State Board of Education, 2000). Appendix E presents the number and percent of responses for each of the standards. As can be seen from the appendix table, some state standards were considered to be directly measured whereas others were seen to be implicitly measured as judged by a majority of the expert raters. Importantly, all of the standards targeted by our work sample assessment were seen to be directly measured by 75% or more of the panel members (this can be seen by cross-referencing the targeted standards in Appendix A with the state standards in Appendix E).

Finally, we examined the frequency of the teaching behaviors in job performance by asking the panel of expert raters to judge how often they would expect a beginning teacher to engage in each of the tasks required by the work sample during the course of his or her professional practice. Level of frequency was rated on a scale of: 1 = Never; 2 = Less Than Once A Year; 3 = A Few Times A Year; and 4 = A Few Times A Week. Appendix F presents the number and percentage of raters for each component of the teacher work sample by frequency level. As can be seen from the appendix table, a majority (68% or more) of the raters indicated a high frequency of a few times a week for each of the work sample components. This results supports the frequency criteria for the content representativeness of our teacher work samples.

Impact on Student Learning

Additional analyses focused on the quality of sources of evidence for student learning. Partial evidence of the impact of teacher performance on K-12 student learning is reflected in the section of our teacher work sample that required teachers to

use assessment data to profile student learning, communicate information about student progress, and plan future instruction based on student learning. In this section of the work sample, teachers must provide an accurate and clear summary of student performance on pre- and post-assessments; evaluate student performance on the achievement targets; use assessment data to draw conclusions about the learning of all students and provide evidence of impacts on student learning; and disaggregate data as needed to inform conclusions about student learning. The key aspect of this section is that to be judged proficient candidates are required to demonstrate an impact on the learning of their students. The first question we considered was whether this section of the work sample could be scored reliably by our raters. The second question we considered was whether performance on this section of the work sample distinguished among the holistic score categorizations of the teachers' performances on the teacher work sample assessment overall.

For the analytic scoring of this Analysis of Learning section of our work samples, we again computed dependability coefficients for absolute decisions using the formulas provided by Crocker and Algina (1986) and Shavelson and Webb (1991). Table 3 presents the analysis of variance for the effect of rater for the three sets of teacher work samples for the analytic scores on this section. As was the case for the total analytic scores, for all three sets, the effect of rater was not statistically significant at the $\alpha = .05$ level of significance. Table 4 presents the variance components used in the formulas for computing dependability for each of the three sets of work samples. Each set of work samples was scored by five different raters. The results yielded 5 rater coefficients of dependability of .92, .73 and .92 respectively for the three sets of work samples. Single rater coefficients of dependability for absolute decisions were computed to be .71, .35,

and .70. Adjusting the number of raters included in the formula revealed an acceptable level of dependability of .62 to .88 could be achieved with three raters.

Insert Table 3 about here

Insert Table 4 about here

The association between the average ratings (averaged across five independent raters) of the quality of assessment of student learning and the holistic performance category of the work samples was assessed using chi-square analysis. The result indicated a significant association between analysis of student learning and the holistic score ratings of the teacher work samples, χ^2 (df = 24) = 37.92, p = .035. The degree of association as assessed by Kendall's Tau-b was .66. A higher degree of association might have been attained had the analytic scoring rubric afforded a distinction between performances that merely met the standard and those that exceeded the standard (and thus should be judged exemplary). Nevertheless, our finding suggests the ability to demonstrate analysis of and impact on student learning was an important factor distinguishing the rated proficiency of teacher work samples along a continuum from beginning to exemplary. Hence, to perform well on our teacher work sample overall, the teachers had to be judged to have provided a quality analysis of student learning and to have impacted the learning of their students positively.

Discussion

This study examined the generalizability and validity of teacher work samples for the purpose of documenting teacher education candidates' ability to meet program and state teaching standards and show impact on the learning of the students they teach. This study also established benchmarks for work samples along a continuum of beginning, developing, proficient, and exemplary. Our benchmarking study yielded significant information relative to the ethical issues and shared responsibilities inherent in teacher work sample assessment. This information also begins to address a number of the criticisms of Teacher Work Sample Methodology (see Airasian 1997; Darling-Hammond, 1997; Popham, 1997; Stufflebeam, 1997) as an approach to using student achievement as a measure of teacher performance.

Levels of Competence

If work samples are to provide credible evidence for making judgments about teacher candidates' performance with respect to program standards and state certification requirements, then they must be shown to differentiate levels of competence in accordance with those standards and requirements. Our results have shown teacher work samples can be clearly differentiated into four distinct groups along a developmental continuum from beginning level to highly expert level on the basis of the degree to which candidates have demonstrated their ability to meet standards. We have also shown that holistic judgments of category membership are validly supported by a more analytic rating of each of the targeted standards. Thus, we have established this important first step to the ethical use of teacher work samples for making valid judgments about candidates' performance for these kinds of high-stakes decisions.

Significantly, the highest percentage of the work samples in this study, 37.1%, were judged to be only at the developing level on the continuum, and less than half of the work samples (only 43.9%) were judged to be at the proficient level or better on the continuum. This result is inconsistent with the one reported by McConney, Schalock, and Schalock (1998) for work samples completed at Western Oregon University. McConney et al. (1998) claim "...the opportunity to evaluate unsuccessful work samples completed as a capstone demonstration of proficiency is extremely rare in part because of their timing and in part because ongoing screening of work sample proficiencies prior to the capstone significantly decreases the likelihood of failure" (p. 360). Our finding, in contrast, indicates that when judgments are made by a panel of experts, which includes non-partisan judges, and judgments are made on the basis of a scoring rubric linked to clearly articulated standards, varying degrees of competency can be identified.

Surprisingly, however, our work thus far has not found an association between work sample quality as measured on our holistic rubric and the source of the work samples. Instead, we found different degrees of quality in the production of work samples at all stages of the developmental continuum from novice to highly experienced teachers. It is possible this outcome reflects the reality of individual performance differences among teachers at all levels--an issue that requires further investigation. This finding may also be due in part to the small number of National Board Certified teachers included in our present sample (something we are attempting to remedy in our current work in progress).

It might also be due in part to the fact that the junior level teacher education candidates received concomitant instruction in the very knowledge and skills to be

demonstrated in the work samples over and beyond the guidance they received in following the directions for completion of the work samples. Thus, a number of our teacher education candidates were able to produce work samples that were judged to be proficient or even exemplary because of this extra scaffolding. Consequently, it remains to be seen whether these students would be able to produce such high quality work samples on their own given less guidance and support. This also raises the ethical problem of control over the amount of assistance provided a candidate in preparing a work sample and the circumstances under which work samples should be developed when high stakes decisions are involved. The kind and level of assistance appears to matter to the judgment reached.

Hence, future research should examine the predictive validity of these holistic judgments as teacher education candidates enter the profession and become teachers themselves. This concern for the predictive validity of work sample assessments has also been acknowledged by McConney et al. (1998).

Content Representativeness

One of the primary ethical issues associated with teacher work sample assessment consists of the valid and authentic representation of the complex process of teaching. As noted by Airasian (1997), this issue can only be addressed through systematic studies of both content and construct validity. Our application of Crocker's (1997) content representativeness approach yielded evidence of the alignment of the teacher work sample tasks with national, state, and institutional standards (content validity) and of the coherence between the teacher work sample tasks and the knowledge base on effective teaching (construct validity). However, only as we track our candidates from this benchmarking study through their first years of teaching will we have even basic

data with respect to predictive and consequential validity (Messsick, 1995). More studies such as our benchmarking study must be completed before states and teacher preparation programs can claim that teacher work sample assessment does indeed provide a valid and authentic measure of a teacher's performance. At present, however, our data does support content representativeness aspects of the validity of teacher work samples for their use in high-stakes decisions about the effectiveness of teaching performance.

Generalizability

We believe the reliability of the decision of whether or not to recommend a teacher candidate for program graduation and certification is an important ethical consideration. Western Oregon University has reported agreement between college and school supervisors with respect to a student teachers' performance in the classroom but have not as yet provided interrater reliability coefficients for other aspects of their Teacher Work Sample Methodology (McConney, Schalock & Schalock, 1998). We believe such coefficients are critical if work sample assessments are to be used for individual, program, or other high-stakes decisions. In addition, we believe it essential to use external expert judges not directly involved in candidate supervision to verify the quality of the ratings made. Thus, we applied concepts from Generalizability Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) to assess the consistency of the scores on our analytic scoring rubric made by a panel of expert raters, which included non-partisan raters.

Generalizability Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) provides a summary coefficient reflecting the level of dependability of raters that is similar in interpretation to classical test theory's reliability coefficient. This

analysis not only enabled us to determine how dependable our experts' ratings were for making absolute decisions about student performance, but also provided us with information with which to determine the appropriate number of raters required for making such decisions. In this study, we have established that a panel of five raters, including external non-partisan raters, were able to achieve a high degree of dependability in their ratings of exemplar work samples. Moreover, it appears that an acceptable level of dependability could be achieved with as few as two raters. Together, our results provide preliminary evidence demonstrating teacher work samples can be administered and scored with sufficient inter-rater dependability to be used to make high-stakes decisions regarding the quality of teaching performance.

Achieving high reliability is, of course, also a matter of rater training. This study has resulted in the identification of a set of benchmarked work samples that can now be used for such training. Hence, our current research is focusing on the level of dependability of the ratings of teacher work samples made using both our analytic and holistic rubrics after raters have been trained. Future investigations should also focus on other aspects of score generalizability. One important aspect to consider is the generalizability of performance ratings across different occasions of work sample development by the same teachers or teacher candidates. Another facet that should be considered is the amount of facilitation teachers and teacher candidates receive when developing their work samples. As mentioned previously, this is an important potential source of measurement error.

Efficiency of Scoring

An important consideration in the use of work sample assessments is whether the work samples can be scored with sufficient efficiency to make them practical for use as

individual and program assessment measures. In this study, we found the average time to score teacher work samples holistically was about ten minutes and the average time for scoring the exemplar work samples analytically was thirteen and a half minutes. Both time estimates are within a range that makes the use of teacher work sample assessment feasible from a practical standpoint. Although the time estimates for analytic scoring were based on exemplar work samples only, our estimates were also from raters who were inexperienced and who had not yet been trained using any exemplars. It is very likely that raters will become more efficient in their time spent rating given both practice and training. Hence, our time estimates may be close enough to reality to draw some tentative conclusions about scoring efficiency. Based on our estimates, we believe a large number of teacher work samples can be scored in a relatively short and reasonable period of time. Other programs can use this data to begin to consider the feasibility of the use of teacher work sample assessments in their own programs.

Impact on Student Learning

An important aspect of our development of teacher work samples has been our effort to link in a defensible way the assessment of teacher performance to the learning of the students they teach. Early in our implementation of teacher work sample assessment we tried the Index of Pupil Growth (Schalock, Schalock & Girod, 1997) developed at Western Oregon University. The Index of Pupil Growth is a direct measure of the learning gains of students in terms of gain scores (Schalock, Schalock & Girod, 1997). The work at Western Oregon has focused on this measure as an indication of the quality of teaching performance. Unfortunately, we found in our pilot work samples that efforts to have our candidates use gain scores as measures of the

learning gains of their students had a negative impact on the significance of the learning goals and the quality and types of assessments candidates employed in their work samples. Use of the index encouraged our candidates to set low-level, non-significant learning goals and to use objective tests rather than other forms of assessment to evaluate student learning. By using the index, we found that we discouraged the very instructional and assessment practices we sought to develop in our candidates. As a result, we quickly abandoned use of the Index of Pupil Growth and began the difficult process of identifying a defensible and credible approach for representing the quality of teaching performance as a function of the learning of their students.

Rather than attempt to measure student learning directly by a single index, our approach has been to set specific criteria for quality teaching performance that takes into consideration the significance of the learning goals, quality of the assessments, and student performance relative to the chosen learning goals. Hence, student learning is addressed by building explicit criteria relative to these factors into our scoring rubrics. Thus, for example, to be judged competent, teachers must provide credible evidence in their work samples that they are able to develop quality pre- and post-assessments of student learning aligned with their achievement targets; are able to disaggregate assessment data on the pre- and post-assessments to profile student accomplishment of the achievement targets; are able to assess the impacts of their instruction on the learning of all students; and are able to communicate information clearly and accurately about student progress. The quality and strength of the evidence determines the rating the work sample receives from our panel of expert raters. We believe this approach avoids many of the pitfalls of efforts to measure student learning on the basis of a single index or test score. However, our approach needs much further work to validate

the judgments of our expert raters with respect to both the quality of the assessments employed by the teachers in their work samples and the quality and quantity of their impacts on student learning.

Nevertheless, in this study, we have demonstrated a significant relationship between holistic performance and the component of the work sample targeting the analysis of student learning. Thus, we have some preliminary evidence which indicates that to be judged competent overall, our teachers and prospective teachers had to provide a quality analysis of student learning and had to demonstrate a positive impact on the learning of their students. While our work in this area is still in its formative stages, this finding indicates that our approach may provide a way to incorporate impacts on student learning into teaching performance assessments that embody national, state, and institutional standards.

Our future work will focus on validating the judgments made on the basis of our scoring rubrics through independent assessments of the impacts of teaching performance on student learning in terms of three dimensions: (1) the quality of the sources of evidence of student learning provided by the candidate in the work samples; (2) the number of students who meet the achievement targets for the instructional sequence; and (3) the number of students who show increased learning (improvement) relative to the achievement targets. We believe these efforts will yield promising information establishing credible links between student learning and assessments of teaching performance.

Shared Responsibility

Through our teacher work sample scoring process and support systems, we have developed a shared responsibility for the preparation of teachers. Professional

education faculty and cooperating teachers work together to create teacher education program course work and field experiences through which our candidates develop the knowledge, skills, and dispositions embodied in our state and institutional standards and targeted in our teacher work sample assessments. The targeted standards, required tasks, and evaluation criteria are clearly communicated and understood by all members of the our professional community, including candidates. In addition, professional education and arts and sciences faculty and practicing educators participate in the scoring of work samples and, as a result, have created a shared knowledge base about assessment and teaching performance. All members of the community -- candidates, university faculty, and practicing educators -- share responsibility for candidate performance and PK-12 student learning.

References

Airasian, P. W. (1997). Oregon teacher work sample methodology: Potential and problems. In J. Millman (Ed.). Grading teachers, grading schools: Is student achievement a valid evaluation measure? (pp. 46-52). Thousand Oakes, CA: Corwin Press.

American Psychological Association. (1985). Standards for educational and psychological testing. Washington, DC: Author

College of Education (1995). Beginning Teacher Core Standards and Indicators. Unpublished Manuscript, Idaho State University, Pocatello, Idaho.

Crocker, L. (1997). Assessing content representativeness of performance assessment exercises. Applied Measurement in Education, 10, 83-95.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Fort Worth, TX: Harcourt, Brace, Jovanovich.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability of scores and profiles. New York: Wiley.

Darling-Hammond, L. (1997). Toward what end? The evaluation of student learning for the improvement of teaching. In J. Millman (Ed.). Grading teachers, grading schools: Is student achievement a valid evaluation measure? (pp. 248-263). Thousand Oakes, CA: Corwin Press.

Idaho State Board of Education (2000). Idaho standards for initial certification of professional school personnel. Boise, ID: Idaho State Department of Education.

McConney, A. A., Schalock, M. D., & Schalock, H. D. (1998). Focusing improvement and quality assurance: Work samples as authentic performance measures

of prospective teachers' effectiveness. Journal of Personnel Evaluation in Education, 11, 343-363.

Messick, S. (1995). Validity of psychological assessments: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50 (9), 741-749.

National Commission on Teaching and America's Future. (1996). What matters most: Teaching and America's Future. New York: Teachers College.

National Council for Accreditation of Teacher Education. (2000). NCATE 2000 Unit Standards. Washington, DC: Author.

Popham, W. J. (1997). The moth and the flame: Student learning as a criterion of instructional competence. In J. Millman (Ed.). Grading teachers, grading schools: Is student achievement a valid evaluation measure? (pp. 264-274). Thousand Oakes, CA: Corwin Press.

Schalock, M. (1998). Accountability, student learning, and the preparation and licensure of teachers: Oregon's teacher work sample methodology. Journal of Personnel Evaluation in Education, 12, 269-285.

Schalock, M., Cowart, B., & Staebler, B. (1993). Teacher productivity revisited: Definition, theory, measurement, and application. Journal of Personnel Evaluation in Education, 7, 179-196.

Schalock, H. D., Schalock, M., & Girod, G. (1997). Teacher work sample methodology as used at Western Oregon State College. In J. Millman (Ed.). Grading teachers, grading schools: Is student achievement a valid evaluation measure? (pp. 15 - 45). Thousand Oakes, CA: Corwin Press.

Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage.

Stufflebeam, D. L. (1997). Oregon teacher work sample methodology: Educational policy review. In J. Millman (Ed.). Grading teachers, grading schools: Is student achievement a valid evaluation measure? (pp. 53-61). Thousand Oakes, CA: Corwin Press.

Table 1

Repeated Measures Analysis of Variance for Effect of Rater on Total Analytic Score Ratings

		F		
Source	df	Set 1	Set 2	Set 3
Rater	4	1.41	.35	2.44
Residual	36	(2.23)	(2.19)	(2.51)

Note. Values enclosed in parentheses represent mean square errors. Set 1 = 10 teacher work samples rated by the same 5 raters. Set 2 = another 10 work samples rated by another 5 raters. Set 3 = final set of 10 work samples rated by another 5 raters.

*$p < .05$

Table 2

Estimates of Variance Components of the Person and Rater Facets for the Total Analytic Score Ratings

Source	Variance Components		
	Set 1	Set 2	Set 3
Person	4.958	3.140	8.660
Rater	.092	-.142	.362
Residual	2.230	2.190	2.510

Table 3

Repeated Measures Analysis of Variance for Effect of Rater on Analysis of Student Learning Ratings

			F	
Source	df	Set 1	Set 2	Set 3
Rater	4	.57	.50	.13
Residual	36	(.23)	(.34)	(.23)

Note. Values enclosed in parentheses represent mean square errors. Set 1 = 10 teacher work samples rated by the same 5 raters. Set 2 = another 10 work samples rated by another 5 raters. Set 3 = final set of 10 work samples rated by another 5 raters.

*$p < .05$

Table 4

Estimates of Variance Components of the Person and Rater Facets for the Analysis of Student Learning Score Ratings

Source	Variance Components		
	Set 1	Set 2	Set 3
Person	.530	.174	.496
Rater	-.010	-.017	-.020

Appendix A

Teacher Work Sample
Assessment Standards And Indicators

Learning-Teaching Context

The teacher uses information about the learning-teaching context and student individual differences to plan instruction and assessment.

Identifies and describes characteristics of the school, classroom, and students; relates characteristics of the school, classroom, and students to instruction; and adapts instruction and assessment to address factors in the learning-teaching context.

Achievement Targets

The teacher sets important, challenging, varied, and appropriate achievement targets.

Provides achievement targets that clearly define what students should know and be able to do; achievement targets are linked to national, state, and local standards and long-term instructional goals; match students' current progress and development; address a variety of learning outcomes; and reflect high expectations for student learning.

Assessment Plan

The teacher uses multiple assessment modes and approaches aligned with achievement targets to assess student learning before, during, and after instruction.

Includes an assessment plan comprised of multiple assessment approaches and modes, including pre-assessments, formative assessments, and post-assessments, that align with achievement targets, and are developmentally appropriate; adapts assessments to accommodate student needs and individual differences; and provides rationales for assessments including validity, useability, and format.

Instructional Sequence

The teacher designs instruction for specific achievement targets, student characteristics and needs, and learning contexts.

Includes learning activities that are aligned with achievement targets and student characteristics and needs; integrates technology into teaching and learning; provides opportunities for collaborations with families; presents accurate and up-to-date content that reflects knowledge of the discipline and modes of inquiry; adapts instruction to accommodate student needs and individual differences.

Analysis of Student Learning

The teacher uses assessment data to profile student learning, communicate information about student progress, and plan future instruction.

Provides and accurate and clear summary of student performance on pre- and post-assessments; uses assessment data to draw conclusions about the learning of ALL students and to evaluate student performance on the achievement targets; disaggregates data as needed to inform conclusions about student learning; provides evidence of the impacts on student learning.

Reflection

The teacher reflects on his or her instruction and student learning in order to improve his or her teaching practice.

Draws conclusions about the extent to which the achievement targets were met and cites evidence to support those conclusions; discusses questions and issues the instructional sequence raised about teaching and students; and reflects on aspects of the instructional sequence that were especially successful or effective and on how the instructional sequence might be taught differently or more effectively.

Appendix B

Teacher Work Sample
Guidelines for Preparation

As a requirement for the Teacher Education Program, you must develop Teacher Work Samples that document your ability to plan, deliver, and assess a standards-based instructional sequence and then reflect on the effects of your instruction on student learning. The Teacher Work Samples will be completed during two of the required teacher education courses: EDUC 309 Planning, Delivery, and Assessment and EDUC 402 Adaptations for Diversity. Through the Teacher Work Sample, you will provide evidence of your performance relative to the following standards:

- The teacher uses information about the learning-teaching context and student individual differences to plan instruction.
- The teacher sets important, challenging, varied, and appropriate achievement targets (i.e., learning goals).
- The teacher uses multiple assessment modes and approaches aligned with achievement targets to assess student learning before, during, and after instruction.
- The teacher designs instruction for specific achievement targets, student characteristics and needs, and learning contexts.
- The teacher adapts instruction and assessment to accommodate student needs and individual differences.
- The teacher uses technology to enhance student learning.
- The teacher collaborates with families to support student learning and development.
- The teacher uses assessment data to profile student learning, communicate information about student progress, and plan future instruction.
- The teacher reflects on his or her instructional practice and on student learning in order to improve his or her teaching practice.

Required Components of the Teacher Work Sample

Your Teacher Work Sample must cover an instructional sequence comprised of at least six learning activities focusing on a concept or set of concepts to be taught over a four-week time period. For your Teacher Work Sample, you will teach lessons and complete a written report. Your report must include the components listed below. Page limitations for each section are noted.

Description and Analysis of the Learning-Teaching Context (2 pages)

In this section of your Teacher Work Sample, you must describe the context in which you teach including the characteristics of the school community, classroom, and students. Before writing this section, you should review class notes and handouts from EDUC 201, EDUC 204, and EDUC 302. This Learning-Teaching Context section of your

Teacher Work Sample must incorporate your knowledge of individual differences, learner characteristics, and environmental factors that impact learning and teaching. *For each factor you describe in this section, you must analyze how that factor impacts instructional planning, delivery, and assessment.*

School characteristics. Provide an overview of important school characteristics including the type of school and grade/subject configuration. You should include any district or state mandates, such as required texts or curricula and content standards, and major characteristics of the local community in which the school is located.

Classroom characteristics. Describe the classroom in which you are completing your pre-internship or internship. You should describe the classroom rules and routines, physical arrangements, grouping patterns, and scheduling that impact learning and teaching in the classroom.

Student characteristics. Describe the students in the classroom including number of students and their ages and gender, cultural and socioeconomic backgrounds, native language(s) and levels of English proficiency, range of abilities, and special needs.

Achievement Targets (1-2 pages)

In this section of your Teacher Work Sample, you must list the achievement targets that will guide the planning, delivery, and assessment of your instructional sequence. The achievement targets for the instructional sequence must clearly define what you expect students to know and be able to do as a result of the instructional sequence. The instructional sequence you use for your Teacher Work Sample must include achievement targets addressing at least three of the following areas: (1) knowledge, (2) reasoning and problem solving, (3) skills, (4) products, and (5) dispositions. Definitions of the areas and example achievement targets are presented in the handout titled "Achievement Targets."

This section of your Teacher Work Sample must also present your rationale for selecting the concept or set of concepts and achievement targets for your instructional sequence. In your rationale, you must identify how your achievement targets (1) relate to the students' current progress and development; (2) align with the classroom teacher's long-range instructional goals; and (3) align with district, state, and national standards.

Assessment Plan (1-3 pages + copies of assessments)

In this section of your Teacher Work Sample, you must design an assessment plan used to monitor student progress toward the achievement target. You must include assessment measures for assessing student performance **before instruction** (pre-assessments), **during instruction** (interim or formative assessments), and **after instruction** (post or summative assessments).

Assessment methods may include paper-and-pencil assessments (i.e., multiple-choice tests and quizzes, essay examinations, written problems, etc.), performance assessments (i.e., reading aloud, communicating conversationally in a second language,

carrying out a specific motor activity in physical education, delivering a speech, etc.), and personal communications (i.e., questions posed and answered during instruction, interviews, conferences, etc.). Your instructional sequence should include a variety of assessment approaches suited for the developmental level of the students and your achievement targets.

The key to writing this section of your Teacher Work Sample is the alignment between your achievement targets and assessment methods. You must construct a table that lists each achievement target, the assessments used to assess student performance relative to the achievement target, a rationale for each assessment that explains why you chose or developed the assessment, and adaptations of the assessments for students with special needs.

Achievement Target	Assessments	Rationale	Adaptations
Achievement Target 1	Pre-Assessment Interim Assessment(s) Post-Assessment	Why you chose or developed each of the assessments for this achievement target.	How you adapted each assessment for students with special needs.
Achievement Target 2			
Achievement Target 3			

Along with the table showing your Assessment Plan for the instructional sequence, you should include copies of the assessments and/or prompts and student directions for the prompts.

Instructional Sequence (12 pages + examples of student work)

This section of your Teacher Work Sample must include individual plans for at least **six** of the learning activities in your instructional sequence. A learning activity can take many forms including, but not limited to, a center, direct whole-group instruction, teacher-directed activity, small-group experience, etc. Your description of each learning activity must include the following items:

1. Content area(s) addressed in the learning activity
2. Grade level(s)
3. Purpose of the learning activity
4. Achievement target(s)
5. Procedures and timeline
6. Materials and resources
7. Adaptations for students with special needs
8. Assessments
9. How integration of technology and outreach to families are included in the learning experience
10. Reflection

The format for writing learning activity plans is attached. With each learning activity plan, you should include samples of student work that represent different levels of performance.

Analysis of Student Learning (1 page + charts or graphs)

In this section of your Teacher Work Sample, you must provide a narrative summary of student learning that occurred as a result of the instructional sequence. You should provide graphs or charts that profile student performance on a pre-assessment and post-assessment used in the instructional sequence. In addition, you should disaggregate data as needed to analyze trends or differences in student learning.

Evaluation and Reflection (2 pages)

For the final section of your Teacher Work Sample, you must write a reflective essay in which you evaluate the effectiveness of your instructional sequence and reflect on your teaching practice and its effects on student learning. You must address the following questions:

- To what extent were the achievement targets for your instructional sequence met? Provide evidence for your response.
- What questions or issues does this instructional sequence reveal about your teaching or the students in your classroom?
- What aspects of your instructional sequence were especially successful or effective? Why?
- How might you teach this instructional sequence differently if you were to do it again? Why?

Format and Organization

Your Teacher Work Sample must include all of the elements listed above and must be word-processed, double-spaced, and error-free. You must adhere to the page limitations for each section. You should provide a Table of Contents that lists the sections of your paper and the page numbers. You must submit your Teacher Work Sample to your course instructor by the deadline date listed in the course syllabus. Your Teacher Work Sample will be evaluated using the attached scoring rubric.

Appendix C

Teacher Work Sample Scoring Rubric

Candidate: _____ Date _____

Evaluator: _____ Level: ___EDUC 309 ___EDUC 402

DIRECTIONS: Using the scale below, please circle the appropriate number to represent the candidate's level of performance on each component of the Teacher Work Sample.

0 = Standard Not Met
Performance fails to provide evidence of meeting the standard for the component of the Teacher work Sample. Performance does not address the indicators of the standard.

1 = Standard Partially Met
Performance provides evidence of partially meeting the standard for the component of the Teacher Work Sample. Performance addresses some of the indicators of the standard.

2 = Standard Met
Performance provides evidence of meeting the standard for the component of the Teacher Work Sample. Performance addresses all of the indicators of the standard.

Learning-Teaching Context	**0**	**1**	**2**
The teacher uses information about the learning-teaching context and student individual differences to plan instruction and assessment. Identifies and describes characteristics of the school, classroom, and students; relates characteristics of the school, classroom, and students to instruction; and adapts instruction and assessment to address factors in the learning-teaching context.			

Achievement Targets	**0**	**1**	**2**
The teacher sets important, challenging, varied, and appropriate achievement targets. Provides achievement targets that clearly define what students should know and be able to do; achievement targets are linked to national, state, and local standards and long-term instructional goals; match students' current progress and development; address a variety of learning outcomes; and reflect high expectations for student learning.			

Assessment Plan	**0**	**1**	**2**
The teacher uses multiple assessment modes and approaches aligned with achievement targets to assess student learning before, during, and after instruction. Includes an assessment plan comprised of multiple assessment approachers and modes, including pre-assessments, formative assessments, and post-assessments, that align with achievement targets, and are developmentally appropriate; adapts assessments to accommodate student needs and individual differences; and provides rationales for assessments including validity, useability, and format.			

Instructional Sequence	**0**	**1**	**2**
The teacher designs instruction for specific achievement targets, student characteristics and needs, and learning contexts. Includes learning activities that are aligned with achievement targets and student characteristics and needs; integrates technology into teaching and learning; provides opportunities for collaborations with families; presents accurate and up-to-date content that reflects knowledge of the discipline and modes of inquiry; adapts instruction to accommodate student needs and individual differences.			

Analysis of Student Learning	**0**	**1**	**2**
The teacher uses assessment data to profile student learning, communicate information about student progress, and plan future instruction. Provides an accurate and clear summary of student performance on pre- and post-assessments; uses assessment data to draw conclusions about the learning of ALL students and to evaluate student performance on the achievement targets; disaggregates data as needed to inform conclusions about student learning; provides evidence of the impacts on student learning.			

Reflection	**0**	**1**	**2**
The teacher reflects on his or her instruction and student learning in order to improve his or her teaching practice. Draws conclusions about the extent to which the achievement targets were met and cites evidence to support those conclusions; discusses questions and issues the instructional sequence raised about teaching and students; and reflects on aspects of the instructional sequence that were especially successful or effective and on how the instructional sequence might be taught differently or more effectively.			

Appendix D

Teacher Work Sample Holistic Scoring Rubric

Teacher Work Sample Holistic Score: Based on the following holistic scoring rubric:

1 Beginning	2 Developing	3 Proficient	4 Exemplary

Beginning

The Beginning performance provides *little or no evidence* of the teacher's ability to plan, deliver, and assess a standards-based instructional sequence and then reflect on his or her instruction and student learning to improve teaching practice.

The Beginning performance provides little or no evidence that the teacher uses information about the learning-teaching context and student individual differences to plan for instruction. When stated, the achievement targets are vague, trivial, inappropriate, or not aligned with national, state, and local standards. The Beginning performance provides little or no evidence that the teacher uses multiple assessment modes and approaches aligned with achievement targets to assess student learning before, during, and after instruction. There is little or no evidence that instruction is designed for specific achievement targets, student characteristics and needs, and learning contexts. Technology is not integrated into teaching and learning and there is little or no collaboration with families to support student learning and development. The Beginning performance provides little or no evidence that the teacher adapts instruction and assessment to accommodate student needs and individual differences. There is little or no evidence that the teacher is able to use assessment data to profile student learning, communicate information about student progress, and plan for future instruction. There is little or no evidence of the impacts on student learning. The Beginning performance provides little or no evidence that the teacher is able to reflect on his or her practice. The reflection is missing or unconnected to instruction and student learning.

Developing

The Developing performance provides *limited evidence* of the teacher's ability to plan, deliver, and assess a standards-based instructional sequence and then reflect on his or her instruction and student learning to improve teaching practice.

The Developing performance provides limited evidence that the teacher uses information about the learning-teaching context and student individual differences to plan for instruction. The achievement targets may be vaguely articulated, of limited significance, or only loosely related to national, state, and local standards. The Developing performance provides limited evidence that the teacher uses multiple assessment modes and approaches aligned with achievement targets to assess student learning before, during, and after instruction. There is limited evidence that instruction is designed for specific achievement targets, student characteristics and needs, and learning context. Technology is minimally integrated into teaching and learning and there is limited collaboration with families to support student learning and development. The Developing performance provides limited evidence that the teacher adapts instruction and assessment to accommodate student needs and individual differences. There is limited evidence that the teacher is able to use assessment data to profile student learning, communicate information about student progress, and plan for future instruction. There is limited evidence of the impacts on student learning. The Developing performance provides limited evidence that the teacher is able to reflect on his or her practice. The teacher is able to describe and analyze his or her practice, but the reflection may be vague, restricted, or focused solely on procedural aspects of teaching.

Proficient

The Proficient performance provides *clear evidence* of the teacher's ability to plan, deliver, and assess a standards-based instructional sequence and then reflect on his or her instruction and student learning to improve teaching practice.

The Proficient performance provides clear evidence that the teacher uses information about the learning-teaching context and student individual differences to plan for instruction. The achievement targets are clear, appropriate, and related to national, state, and local standards. The Proficient performance provides clear evidence that the teacher uses multiple assessment modes and approaches aligned with achievement targets to assess student learning before, during, and after instruction. There is clear evidence that instruction is designed for specific achievement targets, student characteristics and needs, and learning context. Technology is integrated into teaching and learning and efforts are made to collaborate with families to support student learning and development. The Proficient performance provides clear evidence that the teacher adapts instruction and assessment to accommodate student needs and individual differences. There is clear evidence that the teacher is able to use assessment data to profile student learning, communicate information about student progress, and plan for future instruction. There is evidence of the impacts on student learning for the entire class. The Proficient performance provides clear evidence that the teacher is able to reflect on his or her practice. The teacher is able to describe and analyze his or her practice accurately and to reflect on its implications and significance for his or her future teaching.

Exemplary

The Exemplary performance provides *clear, convincing, and consistent* evidence of the teacher's ability to plan, deliver, and assess a standards-based instructional sequence and then reflect on his or her instruction and student learning to improve teaching practice.

The Exemplary performance provides clear, convincing, and consistent evidence that the teacher uses information about the learning-teaching context and student individual differences to plan for instruction. The achievement targets are clear, significant, grounded in national, state, and local standards, and communicate high expectations for all students. The Exemplary performance provides clear, convincing, and consistent evidence that the teacher has a thorough knowledge of individual students and adapts instruction and assessment to meet student needs and individual differences. There is clear, convincing, and consistent evidence that the teacher designs instruction for specific achievement targets, student characteristics and needs, and learning context. Technology is seamlessly integrated into teaching and learning, and the teacher provides multiple opportunities for two-way interactions with families to support student learning and development. Inter- and intradisciplinary connections are made and their use enhances student understanding. There is clear, convincing, and consistent evidence that the teacher is able to accurately describe, analyze, and evaluate each student's performance on the basis of criteria that are known to students and clearly connected to the achievement targets. There is clear, convincing, and consistent evidence of the impacts on student learning for the entire class, subgroups, and individual students. The Exemplary performance provides clear, convincing, and consistent evidence that the teacher is able to describe and analyze his or her practice accurately and to reflect insightfully on its implications and significance for student learning and his or her future teaching.

Appendix E

Number and Percent of Expert Raters Indicating a Match Between Work Sample Assessment and Idaho Core Teacher Standards

Idaho Core Teacher Standards	1 Not at all	2 Implicitly	3 Directly
The teacher understands the central concepts, tools of inquiry, and structures of the discipline(s) taught and creates learning experiences that make these aspects of subject matter meaningful to students.		3 18.8%	13 81.3%
The teacher understands how students learn and develop, and provides opportunities that support their intellectual, social, and personal development.		6 37.5%	10 62.5%
The teacher understands how students differ in their approaches to learning and creates instructional opportunities that are adapted to learners with diverse needs.	1 6.3%	3 18.8%	12 75.0%
The teacher understands and uses a variety of instructional strategies to develop students' critical thinking, problem solving, and performance skills.		4 25%	12 75%
The teacher understands individual and group motivation and behavior and creates a learning environment that encourages positive social interaction, active engagement in learning, and self-motivation.	3 18.8	9 56.3%	4 25%
The teacher uses a variety of communication techniques including verbal, nonverbal, and media to foster inquiry, collaboration, and supportive interaction in and beyond the classroom.		9 56.3%	7 43.8%
The teacher plans and prepares instruction based upon knowledge of subject matter, students, the community, and curriculum goals.		1 6.3%	15 93.8%
The teacher understands, uses, and interprets formal and informal assessment strategies to evaluate and advance student performance and to determine program effectiveness.		2 12.5%	14 87.5%
The teacher is a reflective practitioner who demonstrates a commitment to professional standards and is continuously engaged in purposeful mastery of the art and science of teaching.	1 6.3%	3 18.8%	12 75.0%
The teacher interacts in a professional, effective manner with colleagues, parents, and other members of the community to support students' learning and well-being.	4 25.0%	6 37.5%	6 37.5

Appendix F

Number and Percent of Expert Raters Indicating How Often They Would Expect a Beginning Teacher to Engage in the Teaching Behaviors Required by the Work Sample Assessment

Teaching Behaviors Required by Work Sample Assessment	1 Never	2 Less than once a year	3 A few times a year	4 A few times a week
Use information about the learning-teaching context and student individual differences to plan instruction and assessment.		1 6.3%	3 18.8%	12 75%
Set important, challenging, varied, and appropriate achievement targets.			3 18.8%	13 81.3%
Use multiple assessment modes and approaches aligned with achievement targets to assess student learning before, during, and after instruction.			3 18.8%	13 81.3%
Design instruction for specific achievement targets, student characteristics and needs, and learning contexts.				16 100%
Use assessment data to profile student learning, communicate information about student progress, and plan future instruction.			5 31.3%	11 68.8%
Reflect on his or her instruction and student learning in order to improve his or her teaching.			3 18.8%	13 81.3%